Do Vertical Scales Lead to Sensible Growth Interpretations? Evidence from the Field

Alternate Title:

The Problem with Vertical Scales (It's Not What You Think)

Derek C. Briggs

University of Colorado at Boulder

May 3, 2010

Paper presented at the 2010 Annual Meeting of the American Educational Research Association

Acknowledgments: Thanks to Nathan Dadey for his help gathering and organizing the data used in this paper.

Introduction

This vertical scale permits inferences to be made about student achievement from elementary to high school grades. This vertical scale can be viewed as a developmental continuum. As students develop new capabilities, they move along the continuum, as demonstrated by their scale score. Scale scores are units of a single, equal-interval scale applied across all levels of TerraNova regardless of grade or time of testing. These scores are expressed in numbers that range theoretically from 0 to 999. The equal-interval property of scale scores permits arithmetic functions to be performed using the scale scores. (p. 322)

--Technical Manual for CTB-McGraw Hill's TerraNova Test Battery, 2001.

The developmental scale score is like a ruler that measures growth in reading and math from year to year. Just like height in inches, the student's scores in reading and math are expected to increase each year.

--Newsletter sent to the public from a state board of education[1].

The quotations above are representative of a common assumption about the inherent properties of vertical scales as created through the use of item response theory. It is an assumption made explicitly both by the organizations that craft the scales, and by the stakeholders that make use of them. Interestingly, it is also an

[1] In this article all specific references to states are kept anonymous.

assumption that many—if not most—prominent psychometricians in industry and academia appear not to endorse. This would seem to indicate a gap between theory and practice; a gap recently brought to light in a publication by Dale Ballou (2009). In this paper I investigate the nature of this gap both theoretically and empirically. I do so theoretically by asking the question: if the aim was to create the analog to a ruler for the measurement of mental abilities, on what grounds could a case be made that the endeavor had been a successful one? I do so empirically by comparing growth interpretations from grades 3 through 8 for a subset of 14 states with vertically scaled assessment systems.

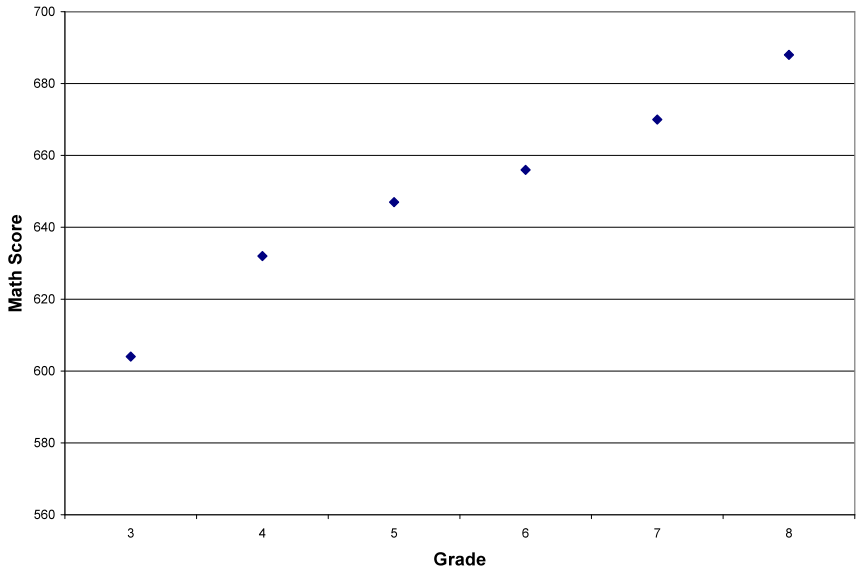
However, before getting to the heart of the matter, I want to begin by posing some premises that I consider relatively uncontroversial, but nonetheless important because I think they establish why this issue is one that is fundamental to test validity and validation, and therefore not just a matter of academic nit-picking.

1. Learning is the principal purpose of sending students to school.
2. Students are tested to find out what they know (and how they know it).
3. We assume that growth in what students know provides a marker that learning has taken place.
4. Teaching should have a positive effect on student growth.

I submit that if you are willing to accept the premises above, then it follows that the creation of a vertical (i.e., developmental) score scale can be plausibly viewed as a precondition to establishing the validity of any standardized assessment. This hinges upon the meaning given to the term "growth". Consider the plot in Figure 1 below. If this is the image that came to mind when you saw the

word growth connected to student learning, then I would argue you are implicitly invoking the concept of a developmental score scale, where the horizontal axis represents the dimension of time (e.g., grade), and a vertical axis represents some construct of knowledge and skills in an academic domain.

Figure 1. An Intuitive Conception of Growth in Student Learning



By invoking the image above it should be clear that one is defining growth in terms of some absolute change in magnitude. In my view it is only in drawing and comparing pictures such as these that we can best evaluate premises 3 and 4 above, along with their relationship to one another.

So at this point I need to lay my cards on the table. In this paper I will argue that the purpose of creating a vertical score scale is to make it possible to compare students *in terms of changes in magnitude.* These changes in magnitude should be in reference to a common unit; that is, the scale should have interval properties. When

this has been accomplished one can claim to *measure* growth in a sense that is both coherent and meaningful. Yet it is probably safe to say that the purpose for creating a vertical scale that I have defined above is not one with which there would be much agreement among the psychometricians responsible for the development of vertical scales for large-scale assessments. I suspect most would argue that vertical scales communicate only ordinal information about student performance. But this points to a problem. If the purpose of vertical scales are used to make absolute statements about growth, then such statements will be entirely equivocal if the scale is ordinal instead of interval. What then, is the point of going through the effort and expense of creating and maintaining a vertical scale?

There are three sections that follow. In the first I present with the existing operational perspective for how one should develop and evaluate a vertical scale and contrast this to what I consider a much stronger approach: the theory of conjoint measurement. In the second, I present results from an analysis of 14 states with existing vertical scales to illustrate how the problems identified with extant theory can cloud subsequent evaluations of practice. In the third section I offer some concluding comments and speculate about future courses of action.

In what follows I will be assuming that the reader has some basic familiarity with the use of item response theory (IRT) modeling in conjunction with a common item test linking design to "calibrate" a vertical score scale. For a recent overview of these methods with many salient references to the existing research literature, see Briggs & Weeks, 2009.

In a recent publication, Ballou (2009) analyzes the theoretical basis behind any claim that a vertical scale created through the use of IRT calibration will have interval properties. Ballou (using a framework attributed to Hand (1996) that was actually first described by Michell (1986)) distinguishes between three theories of measurement that might be invoked to justify an interval interpretation: the classical, operational and representational[2] theories of measurement. He rejects both the classical and operational theories because in the former one simply assumes that the scale is interval a priori, while in the latter one should regard the issue as largely "meaningless" because scale properties are only established with respect to their subsequent use. (I will return to this latter conception momentarily.) In contrast, Ballou considers the view of measurement found in representational theory as providing the only available vehicle for an empirical justification of scale properties. The "vehicle" in this case is primarily the theory of additive conjoint measurement (Luce & Tukey, 1964).

In the most general sense, conjoint additivity implies that two attributes can be scaled such that their additive combination forms a third measure. A classic example of this is of the relationship between force (f), mass (m) and acceleration (a) in Newton's second law of motion after taking logarithms: A = F + M where A = log(a), F = log(f) and M = -log(m). It can be shown that the formulation of the Rasch

[2] Also called axiomatic measurement theory (Krantz et al, 1971). I will use the terms "representational measurement" and "axiomatic measurement" interchangeably throughout.

Model (1960), because it involves the linear and noninteractive combination of person "ability" and item "difficulty" to predict the log odds of a correct response, is one instance of just the sort of situation treated by the theory of conjoint measurement (Brogden, 1977; Perline, Wright & Wainer, 1979; Wright, 1997; Michell, 2008a), since $\log\left[\frac{P(X_{pi}=1)}{P(X_{pi}=0)}\right]=\theta_p - b_i$. The left side of the equation represents the log odds ("logit") of a correct item response and the right side of the equation consists of parameters for a person's ability ("theta", indexed by the subscript *p* for each respondent) and the item's difficulty (b indexed by the subscript *i* for each test item). It is in this sense that one can, on the basis of three key axioms (cancellation, solvability and the Archimedean condition) attempt to justify the logit scale that results from the application of the Rasch model as possessing interval properties[3].

Now by no means do I wish to suggest that it is a trivial matter to satisfy the axioms of conjoint measurement and that if one chooses to apply the Rasch Model to test data, an interval scale is the magic result. The key point is that that theory of conjoint measurement provides a set of criteria that would need to be met before a scale could be described as possessing interval properties. These criteria (i.e., axioms) are internal to the data that is gathered when a standardized test has been administered. To be sure, the axioms would be challenging to satisfy, and only one (cancellation) can be tested directly. But they are the only formal way that can be

[3] A full explication of conjoint measurement and its relationship to the Rasch Model is outside the scope of this paper, but for good presentations see Ballou (2009, 356-364) and Kyngdon (2008a, b).

found in the reseach literture to motivate an empirical distinction between and ordinal and interval score scale.

It has been suggested that the mathematical sophistication necessary to follow the initial presentations of conjoint additivity is the reason that so few psychometricians have taken the time to gain more than a surface level understanding of it. For these and other reasons, as of the early 1990s the psychometrician Norman Cliff (1992) called axiomatic measurement theory the "revolution that never happened." But over the past 20 years a number of very approachable presentations of both the theory of conjoint measurement along with some applications of the theory to real data have made their way into the psychometric research literature (Michell, 1990, 1997; Kyngdon 2008). Furthermore, the mathematics that underlies axiomatic measurement theory is no more daunting than the mathematics found regularly in the pages of *Psychometrika*. So if the purpose of vertical scaling is what I have defined (something that allows for interval comparisons of absolute growth) then we would expect by now to see that some attention to the axioms of conjoint additivity—or at the very least an awareness of them—would be driving recommendations for psychometric practice.

This is unequivocally not the case. The book *Test Equating, Scaling and Linking* by Michael Kolen and Robert Brennan is considered an authoritative source for guidance in methods of developing and maintaining vertcal scales for large-scale assessments. Yet there is no reference at all to conjoint measurement as a basis for scale construction. Instead, the authors assert (p. 332) that "There is no reason to believe, for example, that scores that arise from fitting the Rasch model to

achievement test data are on an interval scale based on the scaling theory of Stevens (1946) and Suppes and Zinnes (1963)." The authors then quote a personal communication from Lord as cited by Angoff (1971) in which the assertion is made that "Any set of measurements can be expressed in terms of a scale with equal units, in some sense, if only we can agree on a definition in operational terms of what is meant by equality." From this Kolen & Brennan conclude that "from the perspective of scaling theory, there is little that can be done to help decide whether one scale is more 'equal interval' than another scale."

Now with all due respect to Kolen & Brennan, two very eminent psychometricians whom I admire, this is simply not true. With the first sentence cited above they dismiss axiomatic measurement theory as a basis for justifying interval scale properties without acknowledging the theoretical basis (i.e., the theory of conjoint measurement) by which this could, in fact, be accomplished if one is interested in quantifying a latent construct. Then in the span of two pages they shift to a purely operational conception of what is meant by the term "measurement" in their invocation of Lord's perspective that anything can be an interval scale "in some sense". Finally, they advance the argument that scaling can only be evaluated with respect to the purposes of a test. At this point, what is being described does not strike me as much of a theory of scaling, because it is not something that can or will ever be falsified. The implicit message from Kolen & Brennan in adopting Lord's operational theory of measurement is that if the purpose of a test is to compare averages, then the scale is de facto interval; if the purpose is to rank students then the scale is de facto ordinal. If this is the theory

driving psychometric practice, then it should be little wonder that psychometricians have lost track of the whole point of creating a vertical scale in the first place.

The operational perspective on scale properties endorsed by Kolen & Brennan appears to derive primarily from two sources: Lord's remarks in an unpublished research bulletin (1950), a one page aside on p. 84 of his 1980 book *Applications of Item Response Theory to Practical Testing Problems*, and a more detailed discussion by Yen (1986) in an article published in the *Journal of Educational Measurement*. Lord had objected to Stevens's (1946; 1951) mapping of scale type (ratio, interval, ordinal, ratio) to permissible statistical procedures, famously claiming "the numbers don't remember where they came from." In the specific case of the scale associated with person parameters (i.e., theta) from IRT models, Lord argued that there was "no obvious reason" to prefer the theta parameterization to some monotonic transformation of theta, using as an example $\theta^* = K\exp(k\theta)$. Now from the perspective of representational measurement theory in general, and the theory of additive conjoint measurement in particular, I think Lord was clearly wrong about this, and it is a point that has been made by Ballou (2009, 362-363) and Briggs & Betebenner (2009). If one's aim in developing a scale is to justify interval properties by satisfying the axioms of conjoint additivity, then any transformation made to theta that does not preserve an additive relationship with item difficulty would violate the assumption of independence necessary for the axiom of cancellation to hold. Only if one is denying the meaningfulness of the distinctions Stevens initially formalized between scale types and permissible

statistics is the contention being made here—that there is no "obvious" reason to prefer one monotonically transformed version of theta to another—a coherent one.

Yen's influential 1986 article adds to the confusion over the scale properties of the theta logit metric when she writes (p. 309)

> IRT produces scale values for items and examinees. These values can be used to make testable, nontrivial predictions about the probability that examinees at different trait (scale) levels will correctly answer each item. When these predictions are accurate, the IRT model produces an equal-interval scale according to the requirements of representational measurement (van der Ven, 1980, p. 255), because only a linear transformation of the item parameter and trait values preserves the model predictions (Allen & Yen, 1979, p. 256).

This statement is confusing for two reasons. First, there is nothing in the citation of van der Ven to support the claim that the IRT model produces an equal interval scale "according to the requirements of representational measurement[4]." Second, the claim that only a linear transformation preserves IRT model predictions conflicts with the perspective just attributed to Lord, a point Yen actually presents later on the same page to seemingly argue that the IRT scale is either inherently ordinal, or that the distinction between ordinal and interval is not something that can be resolved empirically. It is on these grounds that Yen rejects "internal consistency" as a criterion for choosing a test score scale. Instead she focuses the bulk of her article on criteria that are external to the scale itself (e.g., common sense,

[4] van der Ven (1980, p. 255) writes "The level of measurement of the normal ogive model is interval." He makes no attempt to justify this according to representational measurement theory because Lord himself had no interest is such a justification when he proposed a derivation of the IRT model based on the normal ogive (Lord & Novick, 1968, p. 370). Taken in isolation, the normal ogive cdf simply assumes the existence of a continuous variable on an interval scale, so there is no contradiction in van der Ven's statement here unless it is taken out of context.

statistical properties, intended use, etc). There is considerable wisdom is what Yen has to say about the difficulties in using these sorts of subjective criteria to justify scale properties (c.f., Yen, 1986, pp. 311-313). But her position implies that this is the only recourse available, and this seems to be the position that has been driving the psychometric practice of vertical scaling ever since (c.f., Harris, 2007; Patz & Yao, 2007; Young, 2006).

To summarize, the use of the term "interval" in conjunction with IRT scales produced in contemporary vertical scaling practice can only be justified by taking the operational perspective on measurement endorsed by Lord. The seeming contradiction here is that what Lord meant by interval is not what representational measurement theorists, or for that matter, most people mean by interval. This is apparent in the many analogies made to between vertical score scales and the measurement that result from the use of a ruler or thermometer. If one accepts that it is important to distinguish whether a scale has interval as opposed to ordinal properties, then I see no way to avoid the default conclusion that unless evidence can be presented to the contrary, the vertical scales that have resulted from the application of IRT should be assumed to communicate fundamentally ordinal information. Is there any harm in this? I think the answer is yes because it makes growth interpretations based on a vertical scale inherently equivocal. In the next section I demonstrate this using empirical data from 14 states.

The information presented in what follows was gathered between the Fall of 2008 and the Fall of 2009 by visiting the web sites for 24 states reported to have vertically "equated" score scales in the annual "Quality Counts" issue produced by *Education Week* in 2008. For four of these states (Alaska, Minnesota, Mississippi and Rhode Island) no information could be found to support the assertion that these assessment systems contained tests had been vertically scaled. However, two other states (Connecticut and Missouri) were found with vertical scales though they had not been reported as having them by *Education Week*, leaving the total population of states with vertical scales at 22 during the 2007-2008 period. This final number included the following states: Alabama, Arizona, Arkansas, Colorado, Delaware, Florida, Idaho, Illinois, Indiana, Iowa, Mississippi, Missouri, New Mexico, North Carolina, North Dakota, Oregon, South Carolina, South Dakota, Tennessee, West Virginia, Wisconsin, and Wyoming[5]. The most recent and available technical manual and interpretive guide associated with the state's criterion referenced assessment was subsequently retrieved and reviewed. For 14 out of these 22 states, it was possible[6] to locate both the mean scale scores and standard deviations on the state's large-scale math and reading assessments for the grades 3 and 8. In all of these

[5] During the time that this research was being conducted, Texas announced its plans to develop a vertical scale for use in future test administrations. There may be examples of other states contemplating this course of action, but they obviously are outside the scope of the present analysis.

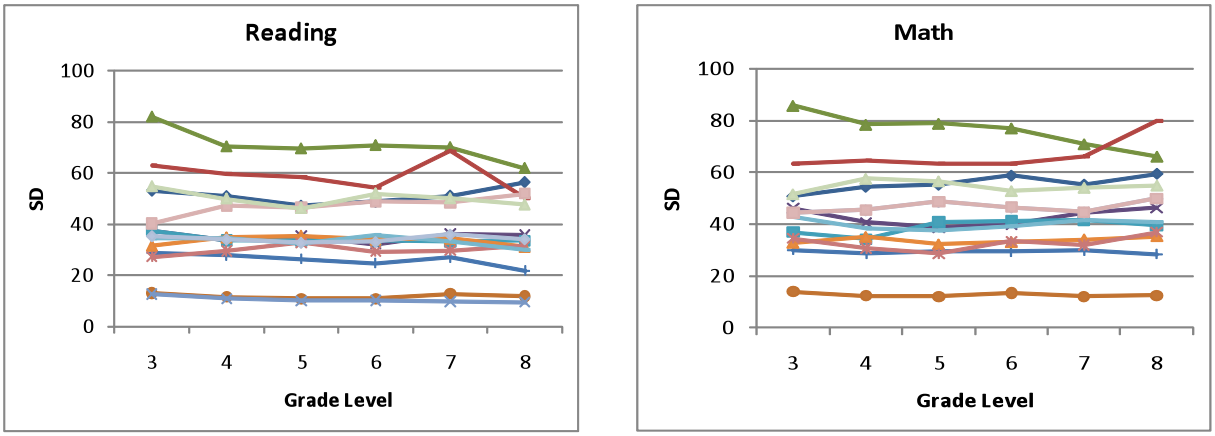
[6] By this I mean that the information was readily available at the state's web site, or it was made available to me upon request. I am still in the process of tracking down this information for the remaining 8 states with vertical scales.

cases IRT methodology was used to create and maintain the vertical scale. When reporting results by state in what follows, the names of specific states are kept anonymous.

Over the past 30 years, vertical score scales have been evaluated on an operational basis, consistent with the scaling "theory" I have sketched out above. What this means in practice is that one examines the empirical patterns of grade by grade means and standard deviations (SDs) after tests have been calibrated onto a common scale. The next step is to determine whether these patterns appear sensible. Along these lines, much of the impetus behind Yen's 1986 publication on vertical scales was the empirical observation of scale shrinkage—that for CTB's test batteries (vertically scaled using the 3PLM) the variability in scale scores decreased, in many cases quite dramatically, as grade level increased. Yen had hypothesized that that this phenomenon could be explained by a violation to the IRT assumption of unidimensionality as items increased in difficulty and complexity across grade levels. In contrast, others had argued that scale shrinkage was merely an artifact of the approach taken for parameter estimation (Camilli, Yamamoto, & Wang, 1993). Another empirical observation that has been frequently made with regard to interpretations of growth along vertical scales is that there is typically a sharp deceleration of growth in the later grades (at the top end of the vertical scale). Some have argued that this is sensible, citing studies that have shown a similar pattern when students at different ages are given the same test and then scores are plotted by age (Martin & Dirir, 2009). Others have argued that this appears symptomatic of "problems" with the use of IRT to establish the scale (Ballou, 2009).

Figure 2 plots the grade by grade SDs in reading and math assessments for the vertical scales of 14 states. (Note that the absolute values on the vertical axis should not be given any interpretation because the SDs do not refer to a common scale metric.) It should be apparent from these plots that among these vertical scales scale shrinkage appears to be the exception rather than the rule. Unless the assessments have become less multidimensional since Yen's research was conducted in the 1980s, this would seem to cast some doubt on the hypothesis that increasing dimensionality over grades is a cause of scale shrinkage.

Figure 2. Assessing Scale Shrinkage in 14 Vertical Scales



Nationally, the two predominant test contractors responsible for the development of state-specific vertical scales between 2007 and 2008 were CTB-McGraw Hill (hereafter "CTB") and Harcourt Educational Measurement (Harcourt)[7]. This is in large part because CTB and Harcourt maintained the commercial test

[7] In 2008, Pearson Educational Measurement acquired Harcourt. So states that had previously contracted with Harcourt became clients Pearson. However, the vertical scale scores reported in this study derive from technical reports that were written by Harcourt staff.

batteries the *TerraNova* and the *Stanford Achievement Test* (SAT10) respectively. Both of these test have vertical scales created using nationally representative samples, a common item nonequivalent groups linking design, and IRT methods to calibrate the vertical scale. For what appears to be a mixture of historical and substantive reasons, the vertical scales produced by CTB tend to involve use of the 3PLM for the calibration of dichotomously scored items and the 2PPCM for the calibration of polytomously scored items. In contrast, vertical scales developed by Harcourt have involved the use of the Rasch Model for dichotomous items in combination with the Partial Credit Model for polytomous items. For most states that contract with CTB, TerraNova items are used to anchor a state-specific test to the Terra Nova vertical scale. Likewise, a similar strategy was often taken among states that contracted with Harcourt in their use of tests items taken from the SAT10 to create a state-specific vertical scale.

Table 1 provides effect size information[8] for states with vertical scales that were (a) developed by CTB or Harcourt, (b) reported (or made available upon request) grade specific scale score means and SDs by grade for any year within the window between 2006 and 2008, and (c) tested students during the spring in a given year. This reduced the available sample of states from the 14 shown in Figure

---

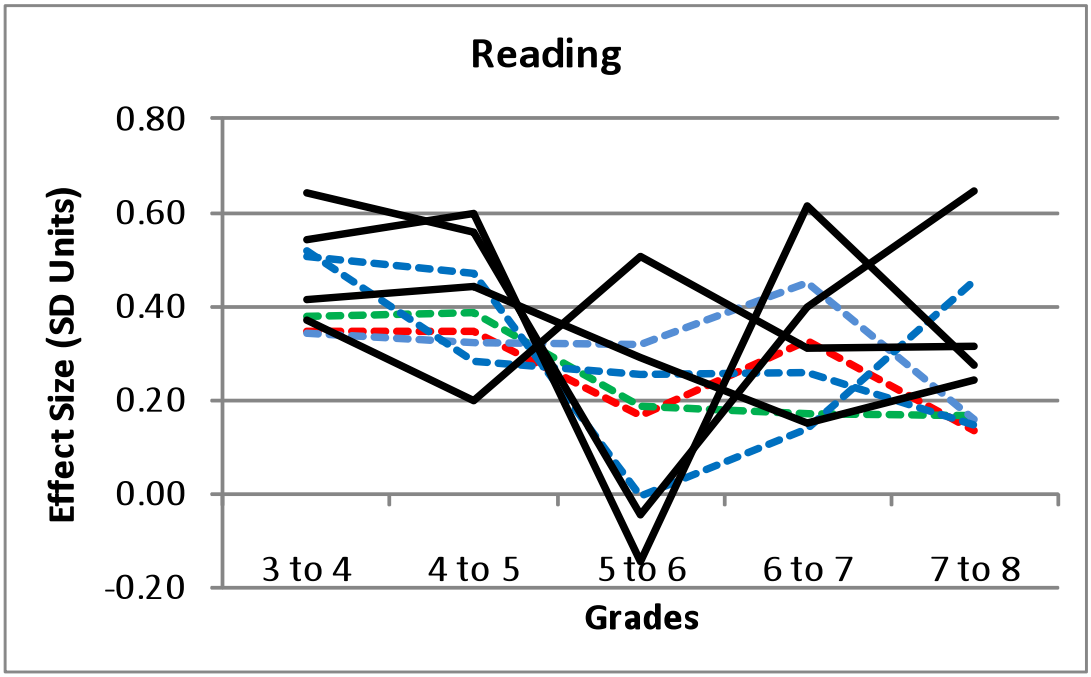
[8] Effect Size = ,where and represent the mean scale scores for the higher and lower grades or years in the scale respectively, and and represent the respective variance for the scores in each grade or year.

2 to 9. The growth across these nine states in math and reading is shown graphically in Figures 3 and 4.

Table 1. Effect Size Growth Patterns for Nine States with Vertical Scales

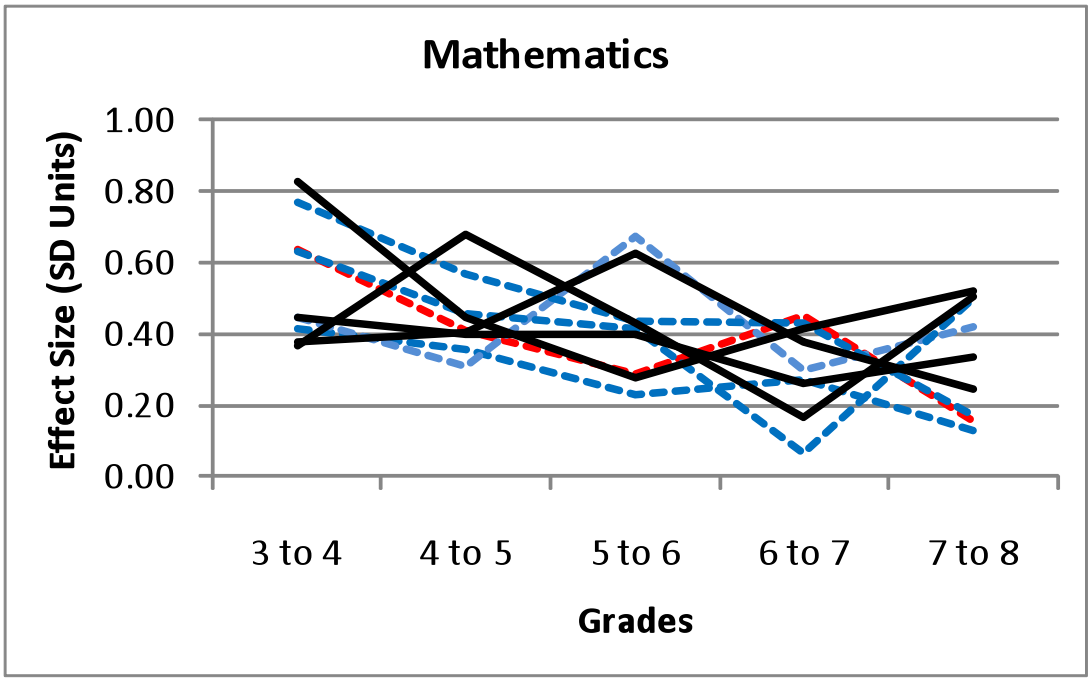
Test Contractor/State		Reading Growth					Math Growth				
		3 to 4	4 to 5	5 to 6	6 to 7	7 to 8	3 to 4	4 to 5	5 to 6	6 to 7	7 to 8
CTB	State A	0.35	0.34	0.17	0.33	0.13	0.64	0.41	0.29	0.45	0.15
	State B	0.38	0.39	0.19	0.17	0.17	0.41	0.36	0.23	0.27	0.13
	State C	0.34	0.32	0.32	0.45	0.16	0.44	0.31	0.67	0.30	0.42
	State D	0.51	0.47	-0.01	0.14	0.46	0.63	0.46	0.42	0.07	0.50
	State E	0.52	0.28	0.25	0.26	0.15	0.77	0.57	0.44	0.43	0.17
Harcourt	State F	0.64	0.56	-0.05	0.40	0.65	0.37	0.68	0.43	0.17	0.51
	State G	0.42	0.44	0.29	0.15	0.24	0.45	0.40	0.40	0.26	0.34
	State H	0.54	0.60	-0.14	0.62	0.27	0.83	0.45	0.28	0.42	0.52
	State I	0.37	0.20	0.51	0.31	0.31	0.38	0.41	0.63	0.38	0.25
	Mean	**0.45**	**0.40**	**0.17**	**0.31**	**0.28**	**0.55**	**0.45**	**0.42**	**0.30**	**0.33**
	Range	**0.30**	**0.40**	**0.65**	**0.48**	**0.51**	**0.46**	**0.37**	**0.44**	**0.38**	**0.39**
	SD	**0.10**	**0.13**	**0.20**	**0.16**	**0.17**	**0.17**	**0.11**	**0.15**	**0.13**	**0.16**

Figure 3. Reading Growth Patterns for 9 States with Vertical Scales



Note: Solid lines = Harcourt (Rasch Model); Dashed lines = CTB (3PL Model)

Figure 4. Math Growth Patterns for 9 States with Vertical Scales



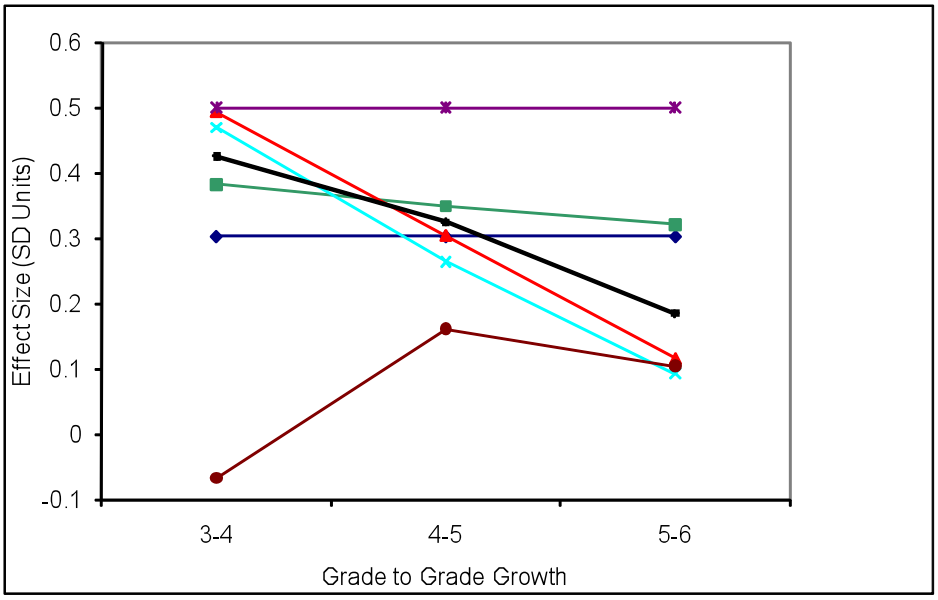
Note: Solid lines = Harcourt (Rasch Model); Dashed lines = CTB (3PL Model)

When examining Figures 3 and 4 what stands out is the variability in these estimates of growth across scales within each of the four adjacent grade combinations (e.g., grade 3 to 4, 4 to 5, etc). Mean growth in student performance for any pair of adjacent grades ranges from a low of .30 SDs (grade 3 to 4 Reading) to a high of .65 SDs (grades 5 to 6 Reading). This variability is much larger for tests of reading than tests of math. With respect to the reading vertical scales, it is also clear that there is greater variability among the four states with scales calibrated using the Rasch family of IRT models (Harcourt, solid dark lines) than there is among the five states using the 3PLM/2PPCM combination (CTB, dashed light lines). For math tests, one sees a trend from grades 3 through 8 consistent with that of decelerating growth—in general, effect sizes decrease in magnitude with increasing

grade. There is no such discernible trend for the reading tests. Numerical summaries of the results in Figures 3 and 4 disaggregated by test contractor are provided in Table 2 below.

One might be tempted to look for other factors germane to the development of a vertical that might explain the variability in growth from state to state beyond the choice of test contractor and, by extension, choice of IRT model for item calibration. For example, perhaps it is the number and choice of common items on tests between adjacent grades that lead to variability in subsequent growth interpretations? Or perhaps the answer can be found in differences in the triangulation of content standards, curricular emphasis, and test design from state to state? Can differences in growth be explained by something schools within a given state are doing to effect learning? One might ask certainly ask these questions, but I suspect the answers may ultimately provide only fool's gold because there is no reason to believe that these vertical scales provide anything beyond an ordinal interpretation. As a result, comparisons of means and SDs such as those made above are largely built upon sand. To illustrate this contention, consider Figure 5 below.

Figure 5. Some Empirical Growth Trajectories



What has been plotted in Figure 5 are seven empirical growth trajectories on the basis of seven distinct vertical scales. Most observers would conclude that these trajectories represent entirely different patterns of growth. In fact, these seven trajectories are based on the exact same underlying set of student item responses from a unique state assessment system. (For details, see Briggs & Betebenner, 2009.) The bold line in black represents effects sizes deriving from this state's publicly reported vertical score scale for reading. All the other lines represent monotonic transformations of the base vertical scale, transformations that are admissible if the underlying scale communicates solely ordinal information. The reader might protest that the practice of monotonically transforming test scores so as to insure that growth takes on some predetermined trajectory and/or predetermined variability is the sort of practice that no commercial test developer would actually support. But they do, and for "theoretical" support they need only

reference Kolen & Brennan (2004), who contend that this is defensible so long as the state has developed a "conceptual definition of growth" and communicated this to the test developer. As an example:

> "The theta scale also can be nonlinearly transformed to provide for growth patterns that *reflect the kind of patterns that are expected* [my emphasis]. Consider a situation in which a test developer believes that the variability of scale scores should increase over grades. If the variability of the theta estimates is not found to increase over grades, a nonlinear transformation of the ability scale might be used that leads to increased variability." (p. 393)

Instances of these sorts of practices were readily found in the technical manuals of two of the states shown above in Figures 3 and 4. The growth trajectories of these two states have been distinguished by red and green colored lines. In the state marked by the red line it was found empirically that the mean scale scores in grade 7 in both reading and math were *lower* than those found in grade 6 *after* the tests had been vertically linked. Rather than report these results, the state—in consultation with its test contractor—decided to adjust the grade 7 scale scores so that the reported mean was that which would have been observed if successive grade means followed a polynomial trend. A similar approach was taken for the reading scale as of grade 8 in the state with the green colored line.

As discussed in the previous section, for the states with vertical scales based on the use of the 3PL and 2PPC models, there is theoretical justification internal to the tests themselves that the resulting scales have interval properties on the basis of the axioms of additive conjoint measurement. This makes any resulting interpretations and comparisons of growth using parametric statistics questionable, for reasons that have just been illustrated. In contrast, it would be possible to make

a case that the states employing the Rasch family of IRT models have vertical scales with interval properties *if* it could be shown that they are probabilistic versions of conjoint measurement. However, this hinges upon careful consideration of whether there is evidence that the Rasch Model fits the data. Showing, for example, that the assumption of non-intersecting item characteristics curves holds empirically is a necessary (but not sufficient) condition for an additive conjoint data structure. Yet among the states that use the Rasch Model and its polytomous extensions to create their vertical scales such evidence is either entirely absent or woefully inadequate.

To illustrate this, consider the case of two states, G and H. In the 2007 technical report for state G's large-scale assessment system, the following claim was made with respect to the fit of the Rasch model: "The statistical fit of the Rasch model to the [*name of test removed*] multiple-choice tests has been previously examined and found to be satisfactory." Having read this I next looked for the results from this "previous examination" in the 2006 technical report. Instead, I found the exact same sentence referring to a previous examination. In fact, the exact same sentence is used in state G's annual technical report dating back 8 years to the version from the year 2000. Prior to that, in the 1999 report, no mention is made of model fit at all. And no criteria was provided in any of these reports for what would constitute fit considered "satisfactory."

By contrast, state H does provide tables of fit statistics for the items used on its various content assessments in its technical reports. Here is how it is suggested that these statistics be interpreted:

> Infit is a statistic that assesses the fit of the observed data to the Rasch IRT model with respect to the parameters that were estimated for that item. Essentially, it answers the question, "How closely does the observed data hold to the values that are predicted by the model?" The infit statistic is sensitive to unexpected responses for examinees with abilities near to the difficulty of the item. Its expected value is 1.0; values greater than 1.5 indicate that the data contains unexpected response patterns.

This proposed interpretation is common among states that use the Rasch Model, and it is usually followed by a table showing that a vast majority of test items have infit statistics within the "acceptable" range between .5 and 1.5. But as a recent paper by Wu & Adams (in press) makes clear, such interpretations are misguided. First, the infit statistic does not answer the question of "how closely the observed values hold to the values predicted by the model." Rather, the statistic gives an indication of the effect size for misfit when there is evidence that the assumption of common slopes among item characteristic curves has been violated. Second, the commonly used "rule of thumb" interval from .5 and 1.5 was derived from a simulation study with a sample size of 100 respondents. If the simulation were to be done with sample sizes in the hundred thousands or millions (as will be the case when using the data for a state), the sampling distribution of the infit statistic under the null hypothesis of equal ICC slopes would be very tightly packed around the expected value of 1. Given this, observing a fit statistic of say, 1.2 or .8, which is currently regarded as acceptable, is in fact a strong indication of a misfitting item. Given this, there is also no good reason to regard the vertical scales developed in states using the Rasch Model as having anything beyond ordinal properties.

Discussion

In a series of publications over the past decade Joel Michell (2000, 2004, 2008a) has argued that the field of psychometrics is a "pathological" science using the following line of reasoning[9]:

a) Psychometricians claim to be able to measure psychological attributes.
b) In this, they imply theories presuming that these attributes are quantitative.
c) There is, however, no hard evidence that these attributes are quantitative.
d) So their claim to be able to measure these attributes is at best premature, at worse false.
e) Instead of putting such claims to the empirical test, the field has erected barriers that prevent its members from recognizing (and thereby responding to) b-d.

It is hard to escape the conclusion that contemporary vertical scaling practices support Michell's thesis. The loophole the field of psychometrics has seemingly invoked to escape the implications of the present critique is to deny the classical definition of measurement (consistent with the purpose for vertical scaling I posited at the outset of this paper) as "the discovery or estimation of the ratio of some magnitude of a quantitative attribute to a unit of that magnitude" (Michell, 1990; 1999) by embracing a philosophy of pragmatism over realism. In my view this is the only philosophically coherent rationale for accepting the currently prevalent

[9] In Michell's presentation of this argument, an attribute that is quantitative is one that by definition has interval properties.

(and conveniently all-encompassing) definition of measurement originally advocated by Stevens (1946) as "the assignment of numerals to objects and events according to rule." An example of this can be seen in a 2007 report on vertical scaling commissioned by The Council of Chief State School Officers written by Richard Patz, senior psychometrician at CTB:

> …these statistical models enable us to estimate and assign numerical values for proficiency given an examinee's responses to test items, they are appropriately called measurement models. Scaling refers to the establishment of units for reporting measures of proficiency, and scaling occurs in conjunction with the identification of measurement models.

Patz later concludes

> A measurement scale should have the property that the units of measurement possess the same meaning across the range of the scale. Although true of familiar measures of height and weight, for example, this property is only at best approximated in scales built for measuring latent variables such as proficiency or achievement. A ten point difference in scale score units may mean something different at the low end of the score scale than it does at the middle or high end, for example. This challenge to interpreting changes in scale scores is made more difficult when vertical scales are involved. Growth from 300 to 350 in scale score units may have a different meaning for third graders than for fifth graders, for example. The failure of achievement scales to achieve true "equal-interval" properties suggests that caution and additional validation efforts are appropriate when changes in scale scores become a focus of interest or accountability.

In these passages Patz is both endorsing a definition of measurement consistent with Stevens's operationally motivated conception while at the same time acknowledging that a test score scale with interval properties cannot be justified through the use of IRT methods. What is not at all clear is what he has in mind when he suggests the need for additional validation efforts given that changes in scale

scores are indeed the focus of current (and in all likelihood future) accountability policies. If there are no fixed criteria available to judge the properties of the scale, how can they be validated beyond a surface level appraisal in which policymakers are asked whether the observed growth matches what they expected? And if there is no a priori psychometric rationale to conclude that vertical scales have interval properties, how does this square with the assertions that can presently be found at the web site of Patz's employer?

> The Scale Score is the basic score for TerraNova and other CTB assessments. It is used to derive all the other scores that describe test performance. Scale Scores can be obtained by one of two scoring methods. The first is Item Response Theory (IRT) item-pattern scoring, a procedure offered only by CTB among the major K-12 test publishers. With item-pattern scoring, Scale Scores are derived numerically using all the information contained in a student's pattern of item responses. The second method is number correct scoring. This method converts the number of correct responses (or points earned for constructed-response items) to a Scale Score. For groups of 25 or more students, the item-pattern and number-correct Scale Scores produce equivalent results. Customers can choose to use either scoring method. CTB recommends item-pattern scoring because it provides more accurate results for individual students. Scale Scores are equal-interval scores that can be averaged and used in other statistical analyses.

(Retrieved from http://www.ctb.com/static/about_assessment/popup_faq6.jsp on April 21, 2010)

I am not necessarily opposed to a pragmatic orientation toward the practice of educational measurement. A pragmatist would argue against any notion of absolute truth underlying the conception of a latent construct in IRT models. In this sense "math ability" is only a convenient label for operationalized knowledge and skills in some content domain. For a pragmatist, the proof is in the pudding of

practical consequences. Along these lines Bob Mislevy has argued that the use of the term "measurement model" in psychometric practice should be regarded as metaphorical. But if psychometricians—at least those in positions of leadership in the United States—wish to explicitly embrace pragmatism, they cannot simultaneously appropriate the syntax and semantics of measurement as it is understood in the physical sciences. If the distinction between ordinal and interval is to be regarded as meaningless, then the consumers of psychometric products should be placed under no illusions to the contrary.

There are some rather thorny issues that need to be resolved to reconcile the creation of vertical scales with the current operational perspectives deriving from Lord's imprint that dominate the research literature. First and foremost we need a better answer to the question of why it is a good idea for large-scale assessments to be placed onto a developmental score scale. If the purpose of vertical scaling is different from the one I defined at the outset of this paper, what is the purpose? It should be clear that any answer having to do with growth[10] implicitly brings us back to the intuition of Figure 1, and that intuition is grounded in an assumption of interval scale properties. If the claim is that the purpose is to produce "quasi-interval" scales this just skirts the issue. Finally, the notion that it should be up to consumers to decide upon a conception of growth that must be met by a vertical scale a priori is little more than an invitation for chicanery.

---

[10] This is true even if the argument is advanced that a vertical scale allows for the possibility of direct comparisons among items administered at different grades. It is not just the ranking of items that is of interest put the magnitude of the distance between items on the logit scale.

There is a great need and demand for vertically scaled tests because there is a great desire to make absolute statements about differences in the quantity of what students have learned. Because of this the distinction between qualitative and quantitative differences is critical, and hence a good argument can be made that we need methods that demarcate ordinal from interval. I think the field has much to gain by rising to this challenge. Imagine if tests were to be developed not solely to satisfy a process by standards blueprint, but to produce an empirical relational system that satisfied the demands of conjoint additivity. This would require test developers to have theories about what makes items harder or easier, and to develop tests with more than a criterion of maximizing score reliability in mind. Instead of a situation in which every state can claim to be "measuring" a unique math and reading construct, we might come to appreciate the (novel?) idea that all these assessments should really be communicating the same fundamental thing about the progress students make in learning during their formative years. We really do need a ruler, but the work will be daunting. Are psychometricians prepared to rise to the challenge while showing greater humility about what our methods can and cannot presently accomplish? The first step is to engage in this conversation, and I hope to that end this paper is a step in the right direction.

References

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*(4), 351–383.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. *Some latent trait models and their use in inferring an examinee's ability. Addison-Wesley, Reading, MA*.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge Univ Press.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*(4), 3–14.

Briggs, D. C. & Betebenner, D. (2009) Is Growth in Student Achievement Scale Dependent? Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA, April 14, 2009.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, *42*(4), 631–634.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, *17*(4), 379.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, 186–190.

Angoff, W. H. (1971) Scales, norms and equivalence scores. In R. L Thorndike (Ed.). *Educational measurement*, (2nd ed., 508-600). Washington, DC: American Council on Education.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*(3), 445–492.

Harris, D. (2007). Practical issues in vertical scaling. In N. Dorans, M. Pommerich & P.. Holland (eds) *Linking and aligning scores and scales*, 233–251, Springer.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer Verlag.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, vol. 1: Additive and polynomial representations*. New York: Academic Press

Kyngdon, A. (2008a). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, *18*(1), 89.

Kyngdon, A. (2008b). Conjoint Measurement, Error and the Rasch Model: A Reply to Michell, and Borsboom and Zand Scholten. *Theory & Psychology*, *18*(1), 125.

Kyngdon, A. (2008c). Treating the Pathology of Psychometrics: An Example from the Comprehension of Continuous Prose Text. *Measurement: Interdisciplinary Research and Perspectives*, 6.

Lord, F. M (1950). *Notes on comparable scales for test scores* (Research Bulletin 5048). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. L. Erlbaum Associates.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

Michell, J. (1986a). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*(3), 398–407.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. L. Erlbaum Associates.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge University Press Cambridge, England.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*(5), 639.

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory & Psychology*, *14*(1), 121.

Michell, J. (2008a). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective*, *6*(1), 7–24.

Michell, J. (2008b). Conjoint measurement and the Rasch paradox: A response to Kyngdon. *Theory & Psychology*, *18*(1), 119.

Patz. R. J. (2007) Vertical scaling in standards-based educational assessment and accountability systems. Technical Report written for the Council of Chief State School Officers.

Patz, R. J. & Yao, L. (2007) Methods and models for vertical scaling. In N. Dorans, M. Pommerich & P.. Holland (eds) *Linking and aligning scores and scales*, 253–272, Springer.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237.

Rasch, G. (1960) *Probablistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. *Handbook of experimental psychology*, 1–49.

Suppes, P. & Zinnes, J. L. (1963) Basic measurement theory. In R. D. Luce, R. R., Bush, & E. Galanter (Eds). *Handbook of mathematical psychology.* New York, NY: John Wiley.

van der Ven, A. (1980). *Introduction to scaling*. J. Wiley, Chichester; New York.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *45*, 51–71.

Wu, M. & Adams, R. J. (in press). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299–325.